1.

证明《统计学习方法》习题1.2

通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当损失函数是对数损失函数时，经验风险最小化等价于极大似然估计。

> 解:
>
> 易得, 假设模型是条件概率分布$P(Y|X)$，损失函数是对数损失函数，即
>
> $$L(Y, f(X)) = -logP(Y|X)$$
>
> 则, 其经验风险最小化为:
>
> $$Remp(f) = \min_{f \in \mathcal{F}} -\frac{1}{N}\sum_{i=1}^{N} logP(Y_i|X_i)$$
>
> 而对数似然函数就等于
>
> $$\sum_{i=1}^{n} \log P(Y_i|X_i)$$
>
> 可以看出，这两个函数只相差一个负号和一个常数因子。因此，最小化经验风险等价于最大化对数似然函数，也就是进行极大似然估计。

2.

请证明下述Hoeffding 引理：

Lemma 1. Let X be a random variable with E(X) = 0 and P(X ∈ [a, b]) = 1. Then it holds

$$Eexp(sX) \le exp(\frac{s^2(b-a)^2}{8})$$

> 令:
>
> $$Y = X - E[X]$$

> 则, Y 也是一个有界随机变量，且 $E[Y]=0$.
>
> 令:
>
> $$Z=\frac{Y-a}{b-a}$$
>
> 则, Z 是一个在 $[0,1]$ 上的有界随机变量, 且
>
> $$E[Z]=\frac{E[X]-a}{b-a}$$
>
> 对任意的: $s\in\mathbb{R}$, 我们有
>
> $$E[e^{sY}]=E[e^{s(b-a)Z+sa}]=e^{sa}E[e^{s(b-a)Z}].$$
>
> 利用Jensen不等式, 我们有
>
> $$E[e^{s(b-a)Z}]\leq e^{s(b-a)E[Z]}=e^{s(b-a)\frac{E[X]-a}{b-a}}.$$
>
> 综合上述两式，我们得到
>
> $$E[e^{sY}]\leq e^{sa}e^{s(b-a)\frac{E[X]-a}{b-a}}=e^{s(E[X]-a)}e^{\frac{s^2(b-a)^2}{4}}.$$
>
> 取 $s=\frac{\lambda}{b-a}$, 我们有
>
> $$E[e^{\lambda Y}]\leq e^{\lambda(E[X]-a)/(b-a)}e^{\frac{\lambda^2(b-a)^2}{4(b-a)^2}}=e^{\frac{\lambda^2(b-a)^2}{8}}.$$
>
> Hoeffding引理得证

3.

请列举一个实际中有监督学习的应用，请说明（1）问题背景、（2）因变量和自变量分别是什么，以及（3）通过机器学习建模如何解决该实际问题。

> 一个实际中有监督学习的应用是垃圾邮件检测。这个问题的背景是，网络上有很多不良的或者无关的邮件，会占用用户的时间和空间，影响用户的体验。因此，需要一种方法来自动地识别和过滤这些垃圾邮件。
>
> 在这个应用中，因变量是邮件是否为垃圾邮件，是一个二分类的问题。自变量是邮件的内

容和特征，例如标题、正文、发件人、附件等。通过机器学习建模，可以利用已经标注好的垃圾邮件和正常邮件作为训练数据，来训练一个分类器，例如朴素贝叶斯、支持向量机、决策树等。然后，用这个分类器来对新收到的邮件进行预测，如果预测为垃圾邮件，则将其移动到一个单独的文件夹中，或者直接删除。这样就可以有效地减少用户收到垃圾邮件的概率，提高用户的满意度。

4.

Please read the background and then prove the following results.
Background:

Let y = Ψ(x), where y is an m × 1 vector, and x is an n × 1 vector. Denote

$$\frac{\partial y}{\partial x^\top}=\begin{bmatrix}\frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n}\\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n}\end{bmatrix}$$

Prove the results:

(a)

Let y = Ax, where y is m×1, x is n×1, A is m×n, and A does not depend on x, then

$$\frac{\partial y}{\partial x^\top}=A$$

不妨令向量 A 为:

$$A=\begin{bmatrix}a_{11} & a_{12} & \cdots & a_{1n}\\ a_{21} & a_{22} & \cdots & a_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ a_{m1} & a_{m2} & \cdots & a_{mn}\end{bmatrix}$$

x 为:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

则易得, y 为:

$$y = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{bmatrix}$$

则:

$$\frac{\partial y}{\partial x^\top} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = A$$

可证.

(b)

Let the scalar α be defined by α = y^TAx, where y is m × 1, x is n × 1, A is m × n, and A is independent of x and y, then

$$\frac{\partial \alpha}{\partial x^\top} = x^\top A^\top (\frac{\partial y}{\partial x^\top}) + y^\top A$$

易得, y^T 为一个 $1 \times m$ 的行向量,

则 Ax 为:

$$Ax = \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} A_i x_i$$

其中, A_i 为 A 的列向量,

则, 易得 α 为:

$$\alpha = y^\top \sum_{i=1}^{n} A_i x_i = \sum_{i=1}^{n} y^\top A_i x_i$$

则,

$$\begin{aligned}
\frac{\partial \alpha}{\partial x^\top} = \begin{bmatrix} \frac{\partial \alpha}{\partial x_1} \\ \frac{\partial \alpha}{\partial x_2} \\ \vdots \\ \frac{\partial \alpha}{\partial x_n} \end{bmatrix} = \begin{bmatrix} y^\top A_1 \\ y^\top A_2 \\ \vdots \\ y^\top A_n \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{n} x_i A_i \frac{\partial y_i}{\partial x_1} \\ \sum_{i=1}^{n} x_i A_i \frac{\partial y_i}{\partial x_2} \\ \vdots \\ \sum_{i=1}^{n} x_i A_i \frac{\partial y_i}{\partial x_n} \end{bmatrix} \\
= y^\top A + x^\top A^\top (\frac{\partial y}{\partial x^\top})
\end{aligned}$$

可证.

(c)

For the special case in which the scalar α is given by the quadratic form α = x^TAx where x is n × 1, A is n × n, and A does not depend on x, then

$$\frac{\partial \alpha}{\partial x^\top} = x^\top (A + A^\top)$$

易得:

$$\begin{aligned}
\alpha &= x^\top A x \\
&= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j
\end{aligned}$$

则:

$$
\begin{aligned}
\frac{\partial \alpha}{\partial x^\top} &= \begin{bmatrix} \frac{\partial \alpha}{\partial x_1} \\ \frac{\partial \alpha}{\partial x_2} \\ \vdots \\ \frac{\partial \alpha}{\partial x_n} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^n A_{1j}x_j + \sum_{i=1}^n A_{i1}x_i \\ \sum_{j=1}^n A_{2j}x_j + \sum_{i=1}^n A_{i2}x_i \\ \vdots \\ \sum_{j=1}^n A_{nj}x_j + \sum_{i=1}^n A_{in}x_i \end{bmatrix} \\
&= \begin{bmatrix} \sum_{j=1}^n A_{1j}x_j \\ \sum_{j=1}^n A_{2j}x_j \\ \vdots \\ \sum_{j=1}^n A_{nj}x_j \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^n A_{i1}x_i \\ \sum_{i=1}^n A_{i2}x_i \\ \vdots \\ \sum_{i=1}^n A_{in}x_i \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1n} & A_{2n} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\
&= x^\top (A + A^\top)
\end{aligned}
$$

> 可证.

(d)

Let the scalar α be defined by $\alpha = y^\top Ax$, where y is m × 1, x is n × 1, A is m×n, and both y and x are functions of the vector z, where z is a q × 1 vector and A does not depend on z. Then

$$\frac{\partial \alpha}{\partial z^\top} = x^\top A^\top (\frac{\partial y}{\partial z^\top}) + y^\top A (\frac{\partial x}{\partial z^\top})$$

> 易得:
>
> α 是一个标量, 即易得:

$$\frac{\partial \alpha}{\partial z^\top} = \begin{bmatrix} \frac{\partial \alpha}{\partial z_1} \\ \frac{\partial \alpha}{\partial z_2} \\ \vdots \\ \frac{\partial \alpha}{\partial z_q} \end{bmatrix}$$

而, 易得:

$$\alpha = y^\top A x = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_j y_i$$

则, 带入得:

$$\begin{aligned} \frac{\partial \alpha}{\partial z^\top} &= \begin{bmatrix} \frac{\partial \alpha}{\partial z_1} \\ \frac{\partial \alpha}{\partial z_2} \\ \vdots \\ \frac{\partial \alpha}{\partial z_q} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_j \frac{\partial y_i}{\partial z_1} \\ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_j \frac{\partial y_i}{\partial z_2} \\ \vdots \\ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} x_j \frac{\partial y_i}{\partial z_q} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} y_i \frac{\partial x_j}{\partial z_1} \\ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} y_i \frac{\partial x_j}{\partial z_2} \\ \vdots \\ \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} y_i \frac{\partial x_j}{\partial z_q} \end{bmatrix} \\ &= x^\top A^\top (\frac{\partial y}{\partial z^\top}) + y^\top A (\frac{\partial x}{\partial z^\top}) \end{aligned}$$

可证.

(e)

Let A be a nonsingular, m × m matrix whose elements are functions of the scalar parameter α. Then

$$\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1} \frac{\partial A}{\partial \alpha} A^{-1}$$

$$
\begin{aligned}
\frac{\partial A^{-1}}{\partial \alpha} &= \lim_{\Delta\alpha\to 0}\frac{(A+\Delta A)^{-1}-A^{-1}}{\Delta\alpha} \\
&= \lim_{\Delta\alpha\to 0}\frac{(A+\Delta A)^{-1}AA^{-1}-(A+\Delta A)^{-1}(A+\Delta A)A^{-1}}{\Delta\alpha} \\
&= \lim_{\Delta\alpha\to 0}\frac{(A+\Delta A)^{-1}(-\Delta A)A^{-1}}{\Delta\alpha} \\
&= -A^{-1}\lim_{\Delta\alpha\to 0}\frac{\Delta A}{\Delta\alpha}A^{-1} \\
&= -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1}
\end{aligned}
$$

5.

Please write $\hat{a}$ as the solution of the minimization problem:

$$\min_a ||Xa-y||_2,$$

where X is a n × p matrix, y is a n × 1 vector and a is a p×1 vector. $X^\top X$ is nonsingular.

将目标函数写成矩阵形式：

$$
\begin{aligned}
f(a) &= ||Xa-y||_2^2 \\
&= (y-Xa)^\top(y-Xa) \\
&= y^\top y - y^\top Xa - a^\top X^\top y + b^\top X^\top Xa \\
&= y^\top y - 2y^\top Xa + a^\top X^\top Xa
\end{aligned}
$$

易得可以将去其分为三个部分:

$$\frac{\partial y^\top y}{\partial a} \tag{1}$$

$$2\frac{\partial y^\top Xa}{\partial a} \tag{2}$$

$$\frac{a^\top X^\top Xa}{\partial a} \tag{3}$$

则, 分别求导:

对 (1) 式:

$$\frac{\partial y^\top y}{\partial a} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{p\times 1}$$

对 (2) 式, 由 $4.(a)$ 得:

$$2\frac{\partial y^\top Xa}{\partial a} = 2(X^\top y)$$

对 (3) 式, 由 $4.(c)$ 得:

$$\begin{aligned}\frac{a^\top X^\top Xa}{\partial a} &= (X^\top X + X^\top X)a \\ &= 2X^\top Xa\end{aligned}$$

则:

$$\frac{\partial (y - Xa)^\top (y - Xa)}{\partial a} = -2X^\top y + 2X^\top Xa$$

令其导数为 0, 则:

$$\begin{gathered} 2X^\top Xa - 2X^\top y = 0 \\ \Longrightarrow \hat{a} = (X^\top X)^{-1} X^\top y \end{gathered}$$